Применение комбинаторных оценок вероятности переобучения в простом голосовании пороговых конъюнкций*

Фрей А. И., Ивахненко А. А, Решетняк И. М.
sashafrey@gmail.com, andrej_iv@mail.ru
Москва, Вычислительный центр им. А. А. Дородницына РАН

Данная работа направлена на практическое применение комбинаторных оценок вероятности переобучения [9, 10, 2, 3] для повышения качества логических алгоритмов классификации. Предлагается эффективный метод вычисления предсказанной информативности, максимизация которой улучшает обобщающую способность отдельных логических закономерностей и их линейной композиции.

Combinatorial generalization bounds with application to simple voting of thresholded conjunction rules*

Frey A. I., Ivakhnenko A. A., Reshetnyak I. M.
Dorodnicyn Computing Centre of RAS, Moscow, Russia

We apply tight combinatorial generalization bounds recently obtained in[9, 10, 2, 3] to enhance rule evaluation heuristic in rule-based classifiers. Experiments on 7 data sets from the UCI ML Repository show that combinatorial bound helps to learn more reliable compositions consisting of less overfitted rules.

Логические закономерности

Рассматривается стандартная постановка задачи классификации. Задано множество объектов $\mathbb{X} = (x_i)_{i=1}^{L}$, описанных n действительными признаками, $x_i = (x_i^1, \ldots, x_i^n)$; каждому объекту x_i соответствует ответ y_i из множества $Y = \{-1, 1\}$.

Логическим правилом называется конъюнкция пороговых предикатов (термов) вида

$$r(x_i) \equiv r(x_i; c^1, \ldots, c^n) = \prod_{j \in \omega} \left[x_i^j \lessgtr_j c^j \right], \qquad (1)$$

где $\omega \subseteq \{1, \ldots, n\}$ — подмножество признаков, $\lessgtr_j$ — одна из операций сравнения $\{\leqslant, \geqslant\}$, c^j — порог по j-му признаку. Говорят, что правило r *выделяет* объект x, если $r(x) = 1$.

Логическая закономерность — это правило, выделяющее достаточно много (p) объектов выбранного класса y (положительных примеров) и приемлемо мало (n) объектов всех остальных классов (отрицательных примеров). Для поиска закономерностей класса y по обучающей выборке $X \subset \mathbb{X}$ решается задача двухкритериальной оптимизации:

$$p(r, X) = \sum_{x_i \in X} r(x_i)\,[y_i = y] \to \max_r;$$
$$n(r, X) = \sum_{x_i \in X} r(x_i)\,[y_i \neq y] \to \min_r.$$

Обычно эта задача сводится к максимизации выбранного скалярного критерия информативности $H(p, n)$. В частности, это может быть точный тест Фишера [8], энтропийный критерий, индекс Джини, тест χ^2, тест ω^2 и другие. В обзоре [7] приведено более 20 критериев, но ни один из них не является безусловно лучшим.

Работа выполнена при финансовой поддержке гранта молодым кандидатам наук президента РФ, МК-5422.2012.9.

Для поиска закономерностей применяются методы дискретной оптимизации: жадные алгоритмы с последующий редукцией правил [6], поиск в ширину [4], генетические алгоритмы [12], асимптотически оптимальные алгоритмы [1] и другие.

Выбор функционала информативности и метода его оптимизации является эвристикой.

Переобучение закономерностей

На практике часто приходится наблюдать эффект переобучения закономерностей — на независимой контрольной выборке пропорция числа положительных p' и отрицательных n' примеров, как правило, смещается в нежелательную сторону: $n'/p' > n/p$. Для сокращения переобучения в [3, 11] предлагается использовать функционал *предсказанной информативности*. Это обычный функционал информативности H, в который вместо величин p, n на известной обучающей выборке подставляются оценки соответствующих величин p', n' на неизвестной контрольной выборке,

$$\tilde{H}(p, n) = H(p - \delta', n + \delta''),$$

где δ' и δ'' — поправки на переобучение, получаемые из комбинаторных оценок вероятности переобучения. Преимущество данного подхода в том, что он совместим с любыми функционалами информативности и любыми алгоритмами поиска закономерностей, поэтому его можно встраивать в стандартные библиотеки. Эксперименты на 6 реальных задачах классификации из репозитория UCI показывают, что максимизация предсказанной информативности улучшает обобщающую способность двух типов композиций закономерностей — взвешенного голосования и решающего списка (голосования по старшинству) [11].

В то же время, недостатком предложенного в [11] алгоритма является относительно низкая численная эффективность. Чтобы вычислить вероятность переобучения для заданного набора признаков, приходится перебирать все конъюнкции, находящиеся в некоторой окрестности выбранной оптимальной конъюнкции. При этом размер окрестности увеличивается экспоненциально с ростом числа признаков (ранга конъюнкции). Кроме того, в [11] предполагается, что значения каждого признака попарно различны на объектах выборки.

В данной работе предлагается ряд упрощений, повышающих численную эффективность и расширяющих границы применимости метода максимизации предсказанной информативности. Во-первых, в оценках вероятности переобучения не учитывается связность, что позволяет применять их для признаков любых типов. Во-вторых, вместо полного перебора конъюнкций по специально построенной окрестности применяется сокращённый перебор только по тем конъюнкциям, для которых в процессе поиска закономерностей было вычислено значение информативности. Данный подход проводит к улучшению обобщающей способности логических закономерностей по сравнению с [11].

Оценки вероятности переобучения

Правило r класса $y \in Y$ индуцирует на $\mathbb{X}$ бинарный вектор ошибок $\big(I(r,x_i)\big)_{i=1}^{L}$, где $I(r,x_i) = \big[r(x_i) \neq [y_i{=}y]\big]$ — индикатор ошибки правила r на объекте x_i. Определим *число* и *частоту* ошибок правила r на выборке $X \subseteq \mathbb{X}$:

$$m(r,X) = \sum_{x_i \in X} I(r,x_i); \quad \nu(r,X) = \frac{m(r,X)}{|X|}.$$

Пусть множество объектов $\mathbb{X}$ разбито две непересекающиеся подвыборки: обучающую X длины ℓ и контрольную $\bar{X}$ длины $k = L - \ell$.

Методом обучения называется отображение, которое произвольной обучающей выборке $X \subseteq \mathbb{X}$ ставит в соответствие некоторое правило $r = \mu X$.

Метод обучения μ называется *монотонным*, если $\mu X = \arg\min\limits_{r} K(r,X)$, где критерий $K(r,X)$ — строго монотонная функция вектора ошибок: для любой пары правил r, v и любой выборки $X \subset \mathbb{X}$ если $I(r,x_i) \leqslant I(v,x_i)$ для всех $x_i \in X$ и хотя бы одно из неравенств строгое, то $K(r,X) < K(v,X)$.

Если функция $H(p,n)$ строго монотонно возрастает по p и строго монотонно убывает по n, то критерий $K(r,X) = -H\big(p(r,X), n(r,X)\big)$ является монотонным, а максимизация информативности — монотонным методом обучения [3]. Все используемые на практике критерии обладают свойством монотонности в указанном смысле.

Для произвольного разбиения $X \sqcup \bar{X} = \mathbb{X}$ *переобученностью* правила r называется уклонение частот его ошибок на контроле и на обучении: $\delta(r,X) = \nu(r,\bar{X}) - \nu(r,X)$.

Следуя слабой вероятностной аксиоматике [9], будем полагать, что на множестве $[\mathbb{X}]^\ell$ всех C_L^ℓ разбиений $(X,\bar{X})$ задано равномерное распределение вероятностей. Тогда вероятность переобучения есть доля разбиений, при которых переобученность превышает заданный порог $\varepsilon \in [0,1]$:

$$Q(\varepsilon) = \mathsf{P}\big[\delta(\mu X, X) \geqslant \varepsilon\big].$$

Заметим, что $1 - Q(\varepsilon)$ есть функция распределения случайной величины $\delta(\mu X, X)$, определённой на конечном вероятностном пространстве $\{[\mathbb{X}]^\ell, 2^{[\mathbb{X}]^\ell}, \mathsf{P}\}$, где P — равномерное распределение.

Пусть R — некоторое множество правил. Обозначим через $X_r = \{x \in \mathbb{X}\colon I(r,x) = 1\}$ множество ошибок правила $r \in R$ на выборке $X \subset \mathbb{X}$. Рассмотрим пару правил $r, v \in R$, такую, что $X_v \subset X_r$. Заметим, что правило r может быть выбрано монотонным методом обучения только для тех разбиений $(X,\bar{X})$, где все объекты $\{x\colon x \in \mathbb{X}_r, x \notin \mathbb{X}_v\}$ лежат в контрольной выборке:

$$[\mu X = r] \leqslant [\mathbb{X}_r \backslash \mathbb{X}_v \subset \bar{X}].$$

В терминах метода порождающих и запрещающих множеств [9] множество

$$\mathbb{X}(r) = \bigcup_{v\colon \mathbb{X}_v \subset \mathbb{X}_r} \mathbb{X}_r \backslash \mathbb{X}_v.$$

является запрещающим для правила r. Следовательно, для любого монотонного метода обучения и любой выборки $\mathbb{X}$ справедлива оценка

$$Q(\varepsilon) \leqslant \sum_{r \in R} \frac{C_{L-q}^{\ell}}{C_L^{\ell}} H_{L-q}^{\ell,m-q}\big(s(\varepsilon)\big) \equiv \eta(\varepsilon), \qquad (2)$$

где $q = |\mathbb{X}(r)|$ — *неполноценность* правила r, $m = m(r,\mathbb{X})$.

Пусть $\varepsilon(\eta)$ — функция, обратная к $\eta(\varepsilon)$. Тогда справедливо утверждение, эквивалентное неравенству (2): с вероятностью не менее $1 - \eta$

$$\nu(r,\bar{X}) \leqslant \nu(r,X) + \varepsilon(\eta).$$

Для построения функционала предсказанной информативности нужны аналогичные оценки частоты ошибок первого и второго рода. Введём множества положительных и отрицательных примеров

$$\mathbb{X}' = \{x_i \in \mathbb{X}\colon y_i = y\}; \quad \mathbb{X}'' = \{x_i \in \mathbb{X}\colon y_i \neq y\};$$

и индикаторы ошибки I и II рода, соответственно:

$$I'(r,x) = [r(x_i) = 0][y_i = y];$$
$$I''(r,x) = [r(x_i) = 1][y_i \neq y].$$

Число и частоту ошибок относительно этих индикаторов обозначим через $m'(r,X)$, $m''(r,X)$, $\nu'(r,X)$ и $\nu''(r,X)$ соответственно.

Следующие формулы являются обобщением оценки (2). Для любого монотонного метода обучения справедливы оценки вероятности переобучения по ошибкам первого и второго рода:

$$Q'(\varepsilon) \leqslant \sum_{r\in R} \frac{C_{L-q}^{\ell}}{C_L^{\ell}} H_{L-q}^{\ell,m'-q'}\left(\tfrac{\ell}{L}\,(m'-\varepsilon k)\right) \equiv \eta'(\varepsilon);$$

$$Q''(\varepsilon) \leqslant \sum_{r\in R} \frac{C_{L-q}^{\ell}}{C_L^{\ell}} H_{L-q}^{\ell,m''-q''}\left(\tfrac{\ell}{L}\,(m''-\varepsilon k)\right) \equiv \eta''(\varepsilon);$$

где $q = |\mathbb{X}(r)|$, $q' = |\mathbb{X}(r)\cap\mathbb{X}'|$, $q'' = |\mathbb{X}(r)\cap\mathbb{X}''|$ — неполноценность правила r относительно индикаторов ошибки I, I', I'' соответственно; $m' = m'(r,\mathbb{X}')$ и $m'' = m''(r,\mathbb{X}'')$ — число ошибок r на $\mathbb{X}$ относительно индикаторов ошибки I', I''.

Теперь построим критерий предсказанной информативности для произвольного $H(p,n)$. Обозначим через $\varepsilon'(\eta)$ и $\varepsilon''(\eta)$ функции, обратные к $\eta'(\varepsilon)$ и $\eta''(\varepsilon)$. В новых обозначениях число положительных и отрицательных примеров во всей выборке $\mathbb{X}$ равны, соответственно, $p = |\mathbb{X}'| - m'(r,\mathbb{X})$ и $n = m''(r,\mathbb{X})$ Возьмём в качестве поправок на переобучение медианные оценки частоты ошибок на контроле, получаемые при $\eta = 0.5$:

$$\tilde{H}(p,n) = H\big(p - L\varepsilon'(0.5),\ n + L\varepsilon''(0.5)\big). \qquad (3)$$

Полученная оценка не накладывает никаких ограничений на то, как именно выбирается множество правил R. Оценки расслоения–связности, использованные в [11], довольно жёстко предполагали, что R — это множество всех правил, получаемых при фиксации набора признаков ω и знаков неравенств $\lessgtr_j$ и варьировании порогов c^j. В этом случае максимизация предсказанной информативности $\tilde{H}(p,n)$ может использоваться только в качестве критерия отбора признаков ω.

Теперь же можно ввести более общее представление процесса поиска закономерностей, считая, что он разбит на *стадии*. На каждой стадии просматривается некоторое множество правил R и из них выбирается лучшее. Критерий $\tilde{H}$ предсказывает, какую информативность выбранное правило будет иметь на новых данных. Для этого используется всё множество правил R, учитывается его сложность и расслоение. Таким образом, критерий $\tilde{H}$ позволяет правильно отранжировать правила, полученные на разных стадиях, но не позволяет сделать правильный выбор внутри каждой стадии.

В данной работе для вычисления поправок на переобучение правила r в качестве множества R использовались все правила того же целевого класса, что и r, построенные алгоритмом поиска закономерностей для признаков, входящих в состав r.

Алгоритм 1. ComBoost (Committee Boosting).

Вход: X — обучающая выборка;
T, l_0, l_1 — параметры;
Выход: композиция правил $a_T = (r_1,\ldots,r_T)$.

1: инициализировать выборку X' и отступы: $X' := X$; $M_i := 0$ для всех $i = 1,\ldots,\ell$;
2: **для всех** $t = 1,\ldots,T$
3: обучить правила r_t^y, $y \in Y$ по выборке X';
4: $(r_t, y_t) := \underset{(r_t^y,y):\, y\in Y}{\arg\min} \sum_{x_i\in X'} [a_t(x_i) \neq y_i]$;
5: обновить значения отступов: $M_i := M_i + y_t y_i r_t(x_i)$, $i = 1,\ldots,\ell$;
6: упорядочить выборку X по возрастанию M_i;
7: $X' := \{x_i \in X : \ell_0 < i \leqslant \ell_1\}$.

Алгоритм 2. Усеченный поиск в ширину.

Вход: X — обучающая выборка;
Θ — семейство термов;
M — максимальный ранг конъюнкции;
S_1 — параметр ширины поиска;
Выход: R — набор правил.

1: инициализация: $R := \varnothing$, $R_0 := \{\varnothing\}$;
2: **для** $m = 1,\ldots,M$
3: $R_m := \varnothing$;
4: **для всех** $r \in R_{m-1}$
5: нарастить правило r термами t: $R_m := R_m \bigcup \{r \wedge t : t \in \Theta$ допустим для $r\}$;
6: выбрать в R_m целевые классы;
7: согласно критерию $\tilde{H}$ оставить в R_m не более S_1 лучших правил за каждый класс;
8: сохранить правила: $R := R \cup R_m$;
вернуть R;

Композиция закономерностей

Простое голосование — это один из стандартных способов построения композиции вида

$$a_t(x) = \operatorname{sign}\left(\sum_{r\in R_{+1}} r(x) - \sum_{r\in R_{-1}} r(x)\right), \qquad (4)$$

состоящей из $t = |R_{-1}| + |R_{+1}|$ логических закономерностей, где R_y — множество закономерностей класса y. Для обучения композиции (4) используется комитетный бустинг ComBoost [5]. В отличие от других разновидностей бустинга, он не взвешивает объекты выборки, а только отбирает подвыборки. Поэтому к методу обучения базовых закономерностей применимы комбинаторные оценки переобучения, существующие только для бинарных функций потерь. Другое важное преимущество ComBoost в том, что, благодаря явной оптимизации распределения отступов, он стремится набрать минимальное достаточное число базовых закономерностей.

На шаге 3 Алгоритма 1 для каждого класса y применяется Алгоритм 2 поиска информативных правил, аналогичный алгоритму ТЭМП [4].

На шаге 5 Алгоритма 2 допустимыми для добавления считаются термы, не содержащие признаков, которые уже вошли в правило r.

Результаты экспериментов

В эксперименте на семи реальных задачах классификации из репозитория UCI сравнивались три варианта ComBoost с точным тестом Фишера в качестве критерия информативности:

A: без поправок на переобучение;

B: с поправками по предложенному методу (3);

C: с поправками по эмпирической оценке $Q(\varepsilon)$, вычисляемой методом Монте-Карло по случайному подмножеству разбиений.

Во всех задачах кроме australian варианты B и C дают лучшее качество классификации тестовых данных. Хотя комбинаторные оценки вычисляются неточно, в некоторых случаях вариант B лучше варианта C. На 5 из 7 задач вариант B даёт лучшие результаты, чем предложенный ранее [11], несмотря на то, что он не учитывает эффект связности. Во всех задачах вариант B имеет существенно меньшую переобученность — разность частоты ошибок между тестовой и обучающей выборками.

Выводы

Предложен эффективный метод вычисления предсказанной информативности для поиска логических закономерностей в задачах классификации. Замена обычного критерия информативности на предсказанную информативность может быть выполнена для любого стандартного метода поиска закономерностей, независимо от вида критерия и механизма перебора правил.

Улучшение обобщающей способности достигается благодаря комбинаторным оценкам вероятности переобучения, учитывающим эффект расслоения семейства правил.

Вычислительная эффективность достигается благодаря тому, что, в отличие от предыдущих работ, не производится никакого дополнительного перебора и оценивания правил — оценки вычисляются только по тем правилам, которые уже были построены в процессе перебора.

	ComBoost-C		ComBoost-B		ComBoost-A	
задача	обуч.	тест	обуч.	тест	обуч.	тест
australian	6.8	14.0	9.9	14.9	6.2	**13.8**
echo-card	0.1	2.3	0.2	**0.9**	0.1	2.4
german	13.1	**25.4**	18.3	27.6	12.9	26.0
heart dis.	8.0	18.9	11.1	**18.5**	7.6	19.3
hepatitis	3.0	19.9	7.8	**18.0**	1.8	21.4
labor	0.6	**8.9**	1.1	11.9	0.5	10.9
liver	11.3	**31.4**	33.0	42.7	8.3	32.3

Таблица 1. Средняя частота ошибок (в процентах) на обучающей и тестовой выборке по различным задачам и различным методом контроля переобучения.

Литература

[1] *Дюкова Е. В., Инякин А. С.* Об асимптотически оптимальном построении тупиковых покрытий целочисленной матрицы // Математические вопросы кибернетики. М.: Физматлит. — 2008. — Вып. 17. — С. 247–262.

[2] *Воронцов К. В.* Комбинаторная теория переобучения: результаты, приложения и открытые проблемы. // Математические методы распознавания образов: 15-ая Всеросс. конф.: Докл. — М.: МАКС Пресс, 2011. — С. 40–43.

[3] *Ивахненко А. А., Воронцов К. В.* Критерии информативности пороговых логических правил с поправкой на переобучение порогов. // Математические методы распознавания образов: 15-ая Всеросс. конф.: Докл. — М.: МАКС Пресс, 2011. — С. 48-51.

[4] *Лбов Г. С.* Методы обработки разнотипных экспериментальных данных — Новосибирск: Наука, 1981.

[5] *Маценов А. А.* Комитетный бустинг: минимизация числа базовых алгоритмов при простом голосовании // Математические методы распознавания образов: 13-ая Всеросс. конф.: Докл. — М.: МАКС Пресс, 2007. — С. 180–183.

[6] *Cohen W. W., Singer Y.* A Simple, Fast and Effective Rule Learner // Proc. of the 16 National Conference on Artificial Intelligence, 1999. — Pp. 335–342.

[7] *Fürnkranz J., Flach P. A.* ROC 'n' Rule Learning-Towards a Better Understanding of Covering Algorithms // Machine Learning. — 2005. — Vol. 58, No. 1. — Pp. 39–77.

[8] *Martin J. K.* An exact probability metric for decision tree splitting and stopping // Machine Learning. — 1997. — Vol. 28, No. 2–3. — Pp. 257–291.

[9] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, No. 3. — Pp. 269–285.

[10] *Vorontsov K. V., Ivahnenko A. A., Reshetnyak I. M.* Generalization bound based on the splitting and connectivity graph of the set of classifiers // Pattern Recognition and Image Analysis: new information technologies (PRIA-10), St. Petersburg, Russian Federation, December 5–12, 2010.

[11] *Vorontsov K. V., Ivahnenko A. A.* Tight Combinatorial Generalization Bounds for Threshold Conjunction Rules // 4-th International Conference on Pattern Recognition and Machine Intelligence, June 27 – July 1, 2011. Lecture Notes in Computer Science. Springer-Verlag, 2011. — Pp. 66–73.

[12] *Yankovskaya A. E, Tsoy Y. R.* Selection of optimal set of diagnostic tests with use of evolutionary approach in intelligent systems // 5-th EUSFLAT Conference New Dimensions in Fuzzy Logic and Related Technologies. — Vol. 2. — Ostrava, Chech Republic, 2007. — Pp. 267–270.